Treasure Trove of Mathematics

SMILES CHEMICAL REACTION DATABASE

LINKS

Pricing Information Sheet make_na server
Purchasing Information
The Wolfram Functions Site
Legal © Notice
Biophysics Software
SMILES Reaction Database
ChemAxon
CFTR genomics
Gene Therapy Net
NCBI database
ChemSpider Database CFTR wiki
Genes and Disease
Mathematica 8 Docs

BLOGS

Treasure Trove of Mathematics

BOOK PREVIEWS

The Gamma Function

OUR BOOKSTORES

The Gamma Function

Questions?
Comments?
Email Us

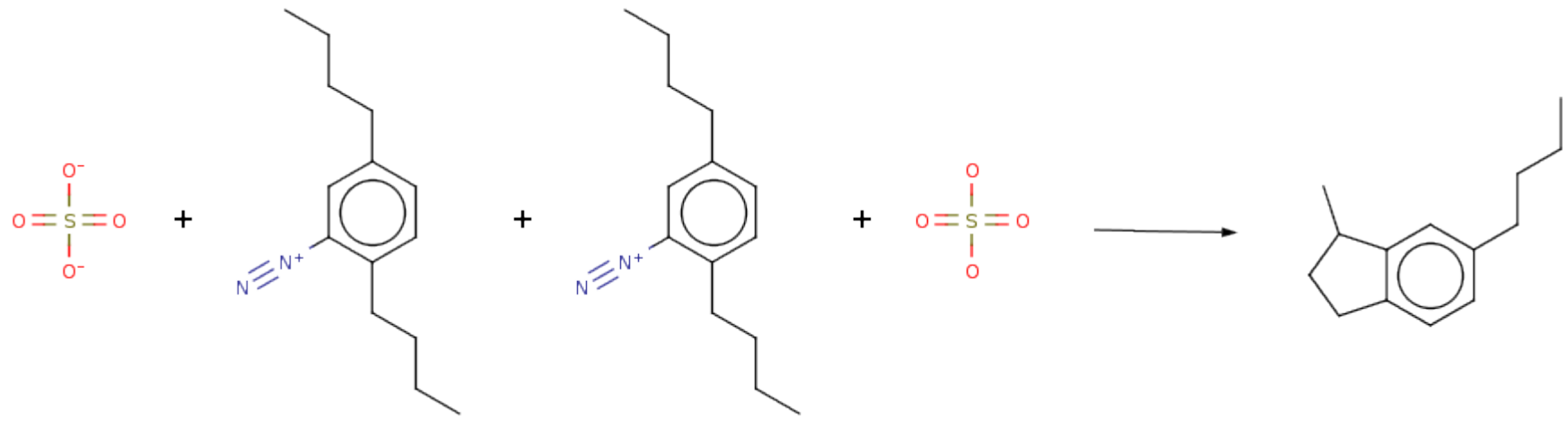
WELCOME

The SMILES Chemical Reaction Database is a set of files containing structural information about pairs of reactant(s) and product(s) of two million different chemical reactions. The simplified molecular-input line-entry system (SMILES) of representing molecular structures is used to represent molecular connectivity and stereochemical relationships as strings of characters, and indeed chemical reactions as well. These SMILES string representations inspired the creation of machine learning computer programs that learn the input/output relationship that exists between *reactant space* and *product space*, using novel string transformation algorithms (implemented within the book *A New Kind of Chemistry © 2012*, scheduled to be released in the Fall of 2012 on Amazon.com, using the Mathematica programming language).

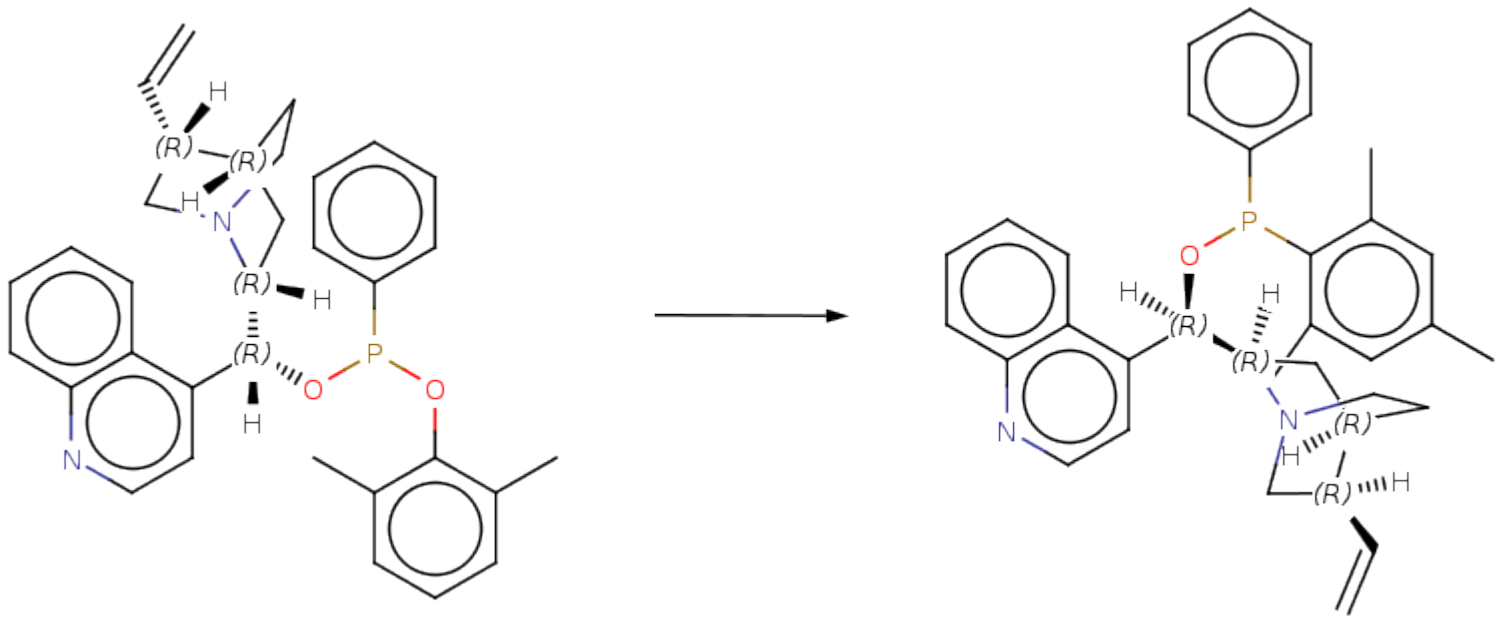
Applications: Chemical Reaction Outcome Prediction, QSARs and Retrosynthetic Analysis.

As a demonstration of the use of SMILES strings to represent the connectivity and steric geometry of chemical structures and reactions, and the utility of the machine learning technique, consider the following two verified results which were correctly predicted by a mathematical model derived from a dataset of 100,000 reactions (of which these two reactions were excluded) possessing reactant profiles (structural and stoichiometric) somewhat similar (very similar cases were excluded for purposes of testing) to each of the novel test cases:

[O-]S([O-])(=O)=O.CCCCc1ccc(CCCC)c(c1)[N+]#N.CCCCc1ccc(CCCC)c(c1)[N+]#N.OS(O)(=O)=O>>CCCCc1ccc2CCC(C)c2c1



[H][C@@](OP(Oc1c(C)cccc1C)c1ccccc1)(c1ccnc2ccccc12)[C@@]1([H])C[C@@]2([H])CCN1C[C@]2([H])C=C>>[H][C@@](OP(c1ccccc1)c1c(C)cc(C)cc1C)(c1ccnc2ccccc12)[C@@]1([H])C[C@@]2([H])CCN1C[C@]2([H])C=C



Of course, the machine learning technique is equally applicable to retrosynthetic analysis – having a target product in mind, one is able to predict the structure of successful starting materials for the prior synthetic step. Many tentative starting materials, or leads, for a synthetic step can be obtained by computing different predictive models, themselves obtained by basing each of the new models on different subsets of the database. Such subsets can be chosen on some selection criteria, or randomly, but in this case each training subset must be entirely composed of reactions having unique sets of reactants to avoid multivalued data.

Reaction prediction is a one-to-one (1:1) relationship whereas retrosynthetic analysis concerns a one-to-many (1: M) relationship. In the case of retrosynthetic analysis, this situation is dealt with by decreasing the size of the training data set to the point where the resulting model makes incorrect suggestions a good fraction of the time. Having not incorporated a significant amount (and possibly type) of knowledge from the database, the model has room to get creative sort of speak. Yet by subsequently running the results through a well-trained reaction prediction model, we borrow back definitiveness, and thereby confirm whether the suggested reactions are feasible or not.

Machine learning of chemical reactions can be distinguished from the more orthodox approaches in three very important ways: First, the work is entirely non-reductionist, explaining chemical reactivity not as the result of the behaviors of the constituent subatomic particles, but rather as the result of higher mathematical conservation laws.

To understand why conservation laws, which represent mathematical symmetries, are used consider any set of non-collinear data points in the Cartesian plane. The number of possible curves which could pass through those data points is infinite. It is highly presumptuous and almost certainly in error to naïvely assume that a smooth curve connecting the data points would represent the intermediate points correctly given an arbitrary curvy data set. Data fitting, which in essence even includes techniques such as neural networks, in and of itself simply cannot be used to generalize data generically. The fact remains that at least one condition must be applied to the curve which would distinguish the curve as *the* solution. And this requires prior knowledge of a model. Data fitting, in any form, is only properly used to tweak the parameters of a model, not to derive a model. This is a very common oversight that plagues much research in the field of computational intelligence.

In this work, we instead search for what is mathematically conserved to within a proportionality factor. The mathematical conservation law H is isomorphic to the linear relationship y=bx, such that $H(m(D_{i,2}))=\mu H(m(D_{i,1}))$ where the $D_{i,j}$ are empirical data points, $\mu \neq 1$ is a proportionality factor and $m(\cdot)$ is the *chemical metric*. Given that the space is discrete and finite, we may legitimately conclude, under the conditions of a sufficiently simple function H, sufficiently large i, and a well-chosen metric, that a mathematical conservation law has been determined, and that the values of the novel points $[H(m(d_{r,1})),H(m(d_{r,2}))]_\mu$ between the empirical points $[H(m(D_{i,1})),H(m(D_{i,2}))]_\mu$ also lie along the straight line connecting the empirical points. The map can then be considered completed and the $d_{r,2}$ can be numerically solved for. The whole point of linearization is that there are aleph-2 possible different curves, a bigger infinity than that of the set of real numbers, aleph-1. But the set of linear rays bound to a particular point is aleph-1, depending only upon the real value of θ.

H is searched for through a process of evolution. Random functional forms are generated, put through rounds of crossover, mutation, simplification and selection. Both task performance and functional simplicity are applied as selective pressures. Simplicity is sought such that we find true conservation functions. An unreasonable effectiveness of the function at task completion is the goal.

When we apply our mathematical model-building technology to the mathematical analogue of the SMILES Reaction Database or any subset thereof, we are applying the very same logic to a subset of *chemical space* – the discrete space of all molecular structures.

The second distinguishing factor is that the high-level mathematical conservation laws we use to predict reactions are based *directly* upon:

- Experimental reaction data – the reaction database stores two million reaction strings.
- Unique string representations of chemical graphs — SMILES.
- Unique, uniformly-sized, order-dependent and reversible mathematical representations of strings as the product of matrix (non-commutative) multiplication using a character-to-matrix substitution.
- Data splicing – defined as data fusion through the discovery of mathematical conservation laws.
    - Evolution of simplest possible function H is key.
    - H is a scalar function, while m is a matrix function.
    - The functional form of H is dependent upon the functional form of m, the value of μ and the $D_{i,k}$.
- Chemical metric – a scalar-valued matrix function based on an advanced theory of prototypicality.

Since the strings are represented by matrices while m(·) is a scalar, we are essentially assigning multidimensional data points to points on the real line. This does *not* lead to the assignment of more than one multidimensional data point to a single point on the real line. In fact the size of the infinity representing all of the points in the plane and the size of the infinity representing all of the points on the real line are the same. Thus unique assignments of all n-dim data points to points on the real line are possible, which is provable. Take a point on a two-dimensional plane (x,y). We can take the digits which we would use to write down x and y and simply interleave them. This interleaving technique results in a real number for every possible point, and no two points on the plane map to the same number. This same argument can be extended to any number of dimensions, as long as we have a finite number of dimensions. The concept of dimension has no effect on the size or cardinality of an infinite space; dimensions are cardinally meaningless. Yet here we are dealing with a discrete hypervolume, a countable infinity if the whole volume is considered, but in this case – a very large finite number. The total number of possible small organic molecules alone that populate 'chemical space' has been estimated to exceed 10^{60}. Reaction space is thus unfathomably large, yet finite.

The third distinguishing factor is that the machine learning technique is both more definitive, more efficient and more capable than the traditional approaches when applied to chemical reaction questions. For example, traditional quantum reactive scattering calculations are typically limited to reactions involving less than six atoms to within any degree of accuracy. Reactive scattering problems involving more than six atoms become effectively intractable due to the combinatoric increases in the number of operations that must be performed on the mathematical objects inherited from quantum theory to get at a reasonable answer.

String transformations have many valuable applications in mathematics and physics as well (for example, the formal technique known as *term rewriting* is used in the field of computer algebra systems).

ABOUT THE SMILES REACTION DATABASE

In 2007, rapid work at TTM began on the assemblage of a human-reviewed chemical reaction database, soon after the development of the supporting image knowledge-extraction and spidering software was finally achieved. The SMILES Reaction Database is now 186.8 MB in size, and it contains two million reactant-product pairs extracted from thousands of respected journals and patents, contained in six files. The reaction data entries in each file of the database occur on consecutive lines of the file, which are delineated by newline characters.

OBTAINING THE SMILES REACTION DATABASE

Legal Notice

Pricing Information Sheet

Purchase an immediate download:

1 2 3 4 5 6 7 8 9 10
(Select purchasing option)

You may download a maximum of three times, so please save your files to a removable disc and store it safely.

Questions or Comments? Email Us